Patent Application
Docket No. 013.0226.01

SYSTEM AND METHOD FOR ARRANGING CONCEPT CLUSTERS IN THEMATIC RELATIONSHIPS IN A TWO-DIMENSIONAL VISUAL DISPLAY SPACE

Field of the Invention

The present invention relates in general to data visualization and, in particular, to a system and method for arranging concept clusters in thematic relationships in a two-dimensional visual display space.

Background of the Invention

Computer-based data visualization involves the generation and presentation of idealized data on a physical output device, such as a cathode ray tube (CRT), liquid crystal diode (LCD) display, printer and the like. Computer systems visualize data through graphical user interfaces (GUIs) which allow intuitive user interaction and high quality presentation of synthesized information.

The importance of effective data visualization has grown in step with advances in computational resources. Faster processors and larger memory sizes have enabled the application of complex visualization techniques to operate in multi-dimensional concept space. As well, the interconnectivity provided by networks, including intranetworks and internetworks, such as the Internet, enable the communication of large volumes of information to a wide-ranging audience. Effective data visualization techniques are needed to interpret information and model content interpretation.

The use of a visualization language can enhance the effectiveness of data visualization by communicating words, images and shapes as a single, integrated unit. Visualization languages help bridge the gap between the natural perception of a physical environment and the artificial modeling of information within the constraints of a computer system. As raw information cannot always be digested as written words, data visualization attempts to complement and, in some instances, supplant the written word for a more intuitive visual presentation drawing on natural cognitive skills.

Effective data visualization is constrained by the physical limits of computer display systems. Two-dimensional and three-dimensional information can be readily displayed. However, *n*-dimensional information in excess of three dimensions must be artificially compressed. Careful use of color, shape and temporal attributes can simulate multiple dimensions, but comprehension and usability become difficult as additional layers of modeling are artificially grafted into the finite bounds of display capabilities.

Thus, mapping multi-dimensional information into a two- or three-dimensional space presents a problem. Physical displays are practically limited to three dimensions. Compressing multi-dimensional information into three dimensions can mislead, for instance, the viewer through an erroneous interpretation of spatial relationships between individual display objects. Other factors further complicate the interpretation and perception of visualized data, based on the Gestalt principles of proximity, similarity, closed region, connectedness, good continuation, and closure, such as described in R.E. Horn, "Visual Language: Global Communication for the 21st Century," Ch. 3, MacroVU Press (1998), the disclosure of which is incorporated by reference.

In particular, the misperception of visualized data can cause a misinterpretation of, for instance, dependent variables as independent and independent variables as dependent. This type of problem occurs, for example, when visualizing clustered data, which presents discrete groupings of data, which are misperceived as being overlaid or overlapping due to the spatial limitations of a three-dimensional space.

Consider, for example, a group of clusters, each cluster visualized in the form of a circle defining a center and a fixed radius. Each cluster is located some distance from a common origin along a vector measured at a fixed angle from a common axis through the common origin. The radii and distances are independent variables relative to the other clusters and the radius is an independent variable relative to the common origin. In this example, each cluster represents a grouping of points corresponding to objects sharing a common set of traits. The radius of the cluster reflects the relative number of objects contained in the grouping. Clusters located along the same vector are similar in theme as are those clusters located on vectors having a small cosine rotation from each other. Thus, the angle relative to a common axis' distance from a common origin is an independent variable with a correlation between the distance and angle reflecting relative similarity of theme. Each radius is an independent variable representative of volume. When displayed, the overlaying or overlapping of clusters could mislead the viewer into perceiving data dependencies where there are none.

Therefore, there is a need for an approach to presenting arbitrarily dimensioned data in a finite-dimensioned display space while preserving independent data relationships. Preferably, such an approach would organize the data according to theme and place thematically-related clusters into linear spatial arrangements to maximize the number of relationships depicted.

There is a further need for an approach to selecting and orienting data clusters to properly visualize independent and dependent variables while compressing thematic relationships for display.

Summary of the Invention

The present invention provides a system and method for organizing and placing groupings of thematically-related clusters in a visual display space. Each cluster size equals the number of concepts (related documents) contained in the cluster. Clusters sharing a common theme are identified. Individual lists of thematically-related clusters are sorted and categorized into sublists of placeable clusters. Anchor points within each sublist are identified. Each anchor point has at least one open edge at which to graft other thematically-related cluster sublists. Cluster sublists are combined at the anchor points to form groupings, which are placed into the visual display space. The most thematically-relevant cluster grouping is placed at the center of the visual display space.

An embodiment provides a system and method for generating a two-dimensional spatial arrangement of a multi-dimensional cluster rendering. A set of clusters is selected from a concept space. The cluster space includes a multiplicity of clusters visualizing document content based on extracted terms. Each cluster in the clusters set shares a common theme including shared terms. An anchor point on at least one such cluster within the clusters set is determined. The anchor point includes at least one open edge formed along a vector defined from the center of the at least one such cluster. The clusters in the clusters set are arranged into an arrangement of adjacent clusters originating from the anchor point at one such open edge.

A further embodiment provides a system and method for arranging concept clusters in thematic relationship in a two-dimensional visual display space. A plurality of stored clusters are selected from a multi-dimensional visualization space sharing a common theme including at least one concept. Each theme logically represents one or more concepts based on terms extracted from a document set. Each cluster not yet grouped from the selected clusters for the shared common theme is combined in order into a list of placeable clusters. Each clusters list is grafted into a grouping of one or more other clusters lists at an anchor point. The anchor point includes an open edge formed along a vector defined from the center of one such cluster in the grouping. The clusters in each other clusters list share at least one concept represented in the shared common theme.

Still other embodiments of the present invention will become readily apparent to those skilled in the art from the following detailed description, wherein is described embodiments of the invention by way of illustrating the best mode contemplated for carrying out the invention. As will be realized, the invention is capable of other and different embodiments and its several details are capable of modifications in various obvious respects, all without departing from the spirit and the scope of the present invention. Accordingly, the drawings and detailed description are to be regarded as illustrative in nature and not as restrictive.

Brief Description of the Drawings

FIGURE 1 is a block diagram showing a system for arranging concept clusters in thematic relationships in a two-dimensional visual display space, in accordance with the present invention.

FIGURE 2 is a graph showing, by way of example, a corpus graph of the frequency of concept occurrences generated by the system of FIGURE 1.

FIGURE 3 is a data representation diagram showing, by way of example, a view of a cluster spine generated by the cluster display system of FIGURE 1.

FIGURES 4A-C are data representation diagrams showing anchor points within cluster spines.

FIGURE 5 is a data representation diagram showing, by way of example, a view of a thematically-related cluster spine grafted onto the cluster spine of FIGURE 4.

FIGURE 6 is a data representation diagram showing, by way of example, a view of singleton clusters and further thematically-related cluster spines grafted onto the cluster spine of FIGURE 5.

FIGURE 7 is a data representation diagram showing, by way of example, a view of a cluster spine of non-circular clusters generated by the cluster display system of FIGURE 1.

FIGURE 8 is a data representation diagram showing, by way of example, a view of a thematically-related cluster spine grafted onto an end-point cluster of the cluster spine of FIGURE 4.

FIGURE 9 is a flow diagram showing a method for arranging concept clusters in thematic relationships in a two-dimensional visual display space, in accordance with the present invention.

FIGURE 10 is a routine for sizing clusters for use in the method of FIGURE 8.

FIGURE 11 is a routine for building sublists of placeable clusters for use in the method of FIGURE 8.

FIGURE 12 is a routine for placing clusters for use in the method of FIGURE 8.

FIGURE 13 is a graph showing, by way of example, an anchor point within a cluster spine generated by the cluster display system of FIGURE 1.

FIGURE 14 is a routine for placing groupers for use in the method of FIGURE 8.

Detailed Description

FIGURE 1 is a block diagram 10 showing a system for arranging concept clusters in thematic relationships in a two-dimensional visual display space, in accordance with the present invention. The system consists of a cluster display system 11, such as implemented on a general-purpose programmed digital computer. The cluster display system 11 is coupled to input devices, including a keyboard 12 and a pointing device 13, such as a mouse, and display 14, including a CRT, LCD display, and the like. As well, a printer (not shown) could function as an alternate display device. The cluster display system 11 includes a processor, memory and persistent storage, such as provided by a storage device 16, within which are stored clusters 17 representing visualized multi-dimensional data. The cluster display system 11 can be interconnected to other computer systems, including clients and servers, over a network 15, such as an intranetwork or internetwork, including the Internet, or various combinations and topologies thereof, as would be recognized by one skilled in the art.

Each cluster 17 represents a grouping of one or more points in a virtualized concept space, as further described below beginning with reference to FIGURE 3. Preferably, the clusters 17 are stored as structured data sorted into an ordered list in ascending or descending order. In the described embodiment, each cluster represents individual concepts and themes extracted from a set of documents 21 and categorized based on, for example, Euclidean distances calculated between each pair of concepts and themes and defined within a pre-specified range of variance, such as described in common-assigned U.S. Patent Application Serial No. 09/944,475, filed August 31, 2001, pending, the disclosure of which is incorporated by reference.

The cluster display system 11 includes three modules: classifier 18, placement 19, and display and visualize 20. The classifier module 18 sorts a list of clusters 17 into either ascending or descending order based cluster sizes. The placement module 19 selects and orients the sized clusters to properly visualize independent and dependent variables while compressing thematic relationships for visual display. The reorient module 19 logically includes a list building submodule for creating sublists of placeable clusters 17, a cluster placement submodule for placing clusters 17 into displayable groupings, known as "groupers," and a grouper placement submodule for placing the groupers within a visual display area. Finally, the display and visualize module 20 performs the actual display of the clusters 17 via the display 14 responsive to commands from the input devices, including keyboard 12 and pointing device 13.

The individual computer systems, including cluster display system 11, are general purpose, programmed digital computing devices consisting of a central processing unit (CPU), random access memory (RAM), non-volatile secondary storage, such as a hard drive or CD ROM drive, network interfaces, and peripheral devices, including user interfacing means, such as a keyboard and display. Program code, including software programs, and data are loaded into the RAM for execution and processing by the CPU and results are generated for display, output, transmittal, or storage.

Each module is a computer program, procedure or module written as source code in a conventional programming language, such as the C++ programming language, and is presented for execution by the CPU as object or byte code, as is known in the art. The various implementations of the source code and object and byte codes can be held on a computer-readable storage medium or embodied on a transmission medium in a carrier wave. The cluster display system 11 operates in accordance with a sequence of process steps, as further described below with reference to FIGURE 8.

FIGURE 2 is a graph showing, by way of example, a corpus graph 30 of the frequency of concept occurrences generated by the system of FIGURE 1. The corpus graph 30 visualizes concepts extracted from a collection of documents 21 (shown in FIGURE 1) represented by weighted clusters of concepts, such as described in commonly-assigned U.S. Patent application Serial No. 09/944,474, filed August 31, 2001, pending, the disclosure of which is incorporated by reference. The *x*-axis 31 defines the individual concepts for all documents 21 and the *y*-axis 32 defines the number of documents 21 referencing each concept. The individual concepts are mapped in order of descending frequency of occurrence 33 to generate a curve 34 representing the latent semantics of the documents set.

A median value 35 is selected and edge conditions 36a-b are established to discriminate between concepts which occur too frequently versus concepts which occur too infrequently. Those documents 21 falling within the edge conditions 36a-b form a subset of documents 21 containing latent concepts. In the described embodiment, the median value 35 is document-type dependent. For efficiency, the upper edge condition 36b is set to 70% and the 64 concepts immediately preceding the upper edge condition 36b are selected, although other forms of threshold discrimination could also be used.

FIGURE 3 is a data representation diagram 40 showing, by way of example, a view 41 of a cluster spine 42 generated by the cluster display system of FIGURE 1. Each cluster in the cluster spine 42, such as endpoint clusters 44 and 46 and midpoint clusters 45, group documents 21 sharing the same themes and falling within the edge conditions 36a-b of the corpus graph 40 (shown in FIGURE 2).

In the described embodiment, cluster size equals the number of concepts contained in the cluster. The cluster spine 41 is built by identifying those clusters 44-46 sharing a common theme. A theme combines two or more concepts 47, which each group terms or phrases (not shown) with common semantic meanings. Terms and phrases are dynamically extracted from a document collection through latent concept evaluation. During cluster spine creation, those clusters 44-46 having available anchor points 48 within each cluster spine 41 are identified for use in grafting other cluster spines sharing thematically-related concepts, as further described below with reference to FIGURE 5.

The cluster spine 41 is placed into a visual display area to generate a two-dimensional spatial arrangement. To represent data inter-relatedness, the clusters

44-46 in each cluster spine 41 are placed along a vector 44 arranged in decreasing cluster size, although other line shapes and cluster orderings can be used.

FIGURES 4A-C are data representation diagrams 50, 60, 65 respectively showing anchor points within cluster spines 51, 61, 66. A cluster having at least one open edge constitutes an anchor point. Referring first to FIGURE 4A, a largest endpoint cluster 52 of a cluster spine 51 functions as an anchor point along each open edge 55a-e. The endpoint cluster 52 contains the largest number of concepts.

An open edge is a point along the edge of a cluster at which another cluster can be adjacently placed. Slight overlap within 20% with other clusters is allowed. An open edge is formed by projecting vectors 54a-e outward from the center 53 of the endpoint cluster 52, preferably at normalized angles. The clusters in the cluster spine 51 are arranged in order of decreasing cluster size.

In the described embodiment, the normalized angles for largest endpoint clusters are at approximately ±60° to minimize interference with other spines while maximizing the degree of interrelatedness between spines. Five open edges 55a-e are available to graft other thematically-related cluster spines. Other evenly divisible angles could be also used. As further described below with reference to FIGURE 5, other thematically-related cluster spines can be grafted to the endpoint cluster 52 at each open edge 55a-e.

Referring next to FIGURE 4B, a smallest endpoint cluster 62 of a cluster spine 61 also functions as an anchor point along each open edge. The endpoint cluster 62 contains the fewest number of concepts. The clusters in the cluster spine 61 are arranged in order of decreasing cluster size. An open edge is formed by projecting vectors 64a-c outward from the center 63 of the endpoint cluster 62, preferably at normalized angles.

In the described embodiment, the normalized angles for smallest endpoint clusters are at approximately ±60°, but only three open edges are available to graft other thematically-related cluster spines. Empirically, limiting the number of available open edges to those facing the direction of decreasing cluster size helps to maximize the interrelatedness of the overall display space.

Referring finally to FIGURE 4C, a midpoint cluster 67 of a cluster spine 61 functions as an anchor point for a cluster spine 66 along each open edge. The midpoint cluster 67 is located intermediate to the clusters in the cluster spine 66 and defines an anchor point along each open edge. An open edge is formed by projecting vectors 69a-b outward from the center 68 of the midpoint cluster 67, preferably at normalized angles. Unlike endpoint clusters 52, 62 the midpoint cluster 67 can only serve as an anchor point along tangential vectors non-coincident to the vector forming the cluster spine 66. Accordingly, endpoint clusters 52,62 include one additional open edge serving as a coincident anchor point.

In the described embodiment, the normalized angles for midpoint clusters are at approximately ±60°, but only two open edges are available to graft other thematically-related cluster spines. Empirically, limiting the number of available open edges to those facing the direction of decreasing cluster size helps to maximize the interrelatedness of the overall display space.

FIGURE 5 is a data representation diagram 70 showing, by way of example, a view 71 of a thematically-related cluster spine 75 grafted onto the cluster spine 42 of FIGURE 3. Each cluster in the cluster spine 75, including endpoint cluster 74 and midpoint cluster 75, share concepts in common with the midpoint cluster 76 of the cluster spine 42. Accordingly, the cluster spine 72 is "grafted" onto the cluster spine 42 at an open edge of an available anchor point on midpoint cluster 76. The combined grafted clusters form a cluster grouping or "grouper" of clusters sharing related or similar themes.

FIGURE 6 is a data representation diagram 80 showing, by way of example, a view 81 of singleton clusters 86 and further thematically-related cluster spines 82 and 84 grafted onto the cluster spine 42 of FIGURE 3. The clusters in the cluster spines 82 and 84 share concepts in common with the clusters of cluster spine 42 and are grafted onto the cluster spine 82 at open edges of available anchor points. Slight overlap 87 between grafted clusters is allowed. In the described embodiment, no more than 20% of a cluster can be covered by overlap. The singleton clusters 86, however, do not thematically relate to the clusters in cluster spines 62, 72, 82, 84 and are therefore grouped as individual clusters in non-relational placements.

FIGURE 7 is a data representation diagram 100 showing, by way of example, a view 101 of a cluster spine 102 of non-circular clusters 104-106 generated by the cluster display system of FIGURE 1. Each cluster in the cluster spine 102, including endpoint clusters 104, 106 and midpoint clusters 105, has a center of mass c_{m} 107a-e and is oriented along a common vector 103.

As described above, with reference to FIGURE 3, each cluster 104-106 represents multi-dimensional data modeled in a two-dimensional visual display space. Each cluster 104-106 is non-circular and defines a convex volume representing data located within the multi-dimensional concept space. The center of mass c_{m} 107a-e for each cluster 104-106 is logically located within the convex volume and is used to determine open edges at each anchor point. A segment is measured from the center of mass c_{m} 107a-e for each cluster 104-106. An open edge is formed at the intersection of the segment and the edge of the non-circular cluster. By way of example, the clusters 104-106 represent non-circular shapes that are convex and respectively comprise a circle, a square, an octagon, a triangle, and an oval, although other forms of convex shapes could also be used, either singly or in combination therewith, as would be recognized by one skilled in the art.

FIGURE 8 is a data representation diagram showing, by way of example, a view of a thematically-related cluster spine grafted onto an end-point cluster of the cluster spine of FIGURE 3.

Further thematically-related cluster spines 112, 114, 116, 118 are grafted into the cluster spine 62. The cluster spines 112, 114, 118 are grafted into the largest endpoint cluster of the cluster spine 62 with the cluster spine 112 oriented along a forward-facing axis 113 and the cluster spine 114 oriented along a backward-facing axis 115. The cluster spine 116 is grafted onto a midpoint cluster of the cluster spine 114 along a backward-facing axis 117. Note the cluster spine 116 has overlap 119 with a cluster in the cluster spine 114.

FIGURE 9 is a flow diagram showing a method for arranging concept clusters in thematic relationships in a two-dimensional visual display space 120, in accordance with the present invention. The method presents arbitrarily dimensioned concept data visualized in a two-dimensional visual display space in a manner that preserves independent data relationships between clusters.

First, individual clusters are sized by number of concepts (related documents) contained in each cluster (block 121), as further described below with reference to FIGURE 10. The sized clusters are then analyzed to find shared terms (block 122). In the described embodiment, those clusters sharing one or more semantically relevant concepts are considered thematically-related.

The lists of shared terms are then sorted into sublists of clusters based on the number of clusters that share each term (block 123). The sublists are arranged in order of decreasing cluster size. Next, lists of placeable clusters are built (block 124), as further described below with reference to FIGURE 10. Each list contains those clusters sharing a common theme and which had not yet been placed in the visual display space. The clusters in each sublist are placed into individual groupings or "groupers" to form cluster spines (block 125), as further described below with reference to FIGURE 11. The method then terminates.

FIGURE 10 is a routine for sizing clusters 130 for use in the method of FIGURE 8. The purpose of this routine is to determine the size of each cluster based on the number of concepts contained in the cluster.

Each cluster is iteratively sized (blocks 131-133) as follows. For each cluster (block 131), the cluster size is set to equal the number of concepts contained in the cluster (block 132). Iterative processing continues (block 133) for each remaining cluster. The groupers are then placed into the visual display space (block 134), as further described below with reference to FIGURE 13. Finally, the placed groupers are displayed (block 135), after which the routine terminates.

FIGURE 11 is a flow diagram showing a routine for building sublists of placeable clusters 140 for use in the method of FIGURE 9. The purpose of this routine is to build sublists of thematically-related clusters to form individual cluster spines. The cluster spines are the building blocks used to form cluster groupings or "groupers."

The sublists are built by iteratively processing each shared concept (blocks 141-150) as follows. For each new shared concept (block 141), a sublist of clusters belonging to the shared concept is built (block 142). A cluster center represents a seed value originating from the shared concept. A seed value typically consists of the core set of concepts, preferably including one or more concepts, which form the basis of the current sublist. Thereafter, each of the clusters is iteratively processed (blocks 143-149) to determine sublist membership, as follows.

For each cluster (block 143), if the cluster does not belong to the current sublist (block 144), that is, the cluster does not share the common concept, the cluster is skipped (block 149). Otherwise, if the cluster has not already been placed in another sublist (block 145), the cluster is added to the current sublist (block 146). Otherwise, if the cluster has been placed (block 145) and has an open edge (block 147), the cluster is marked as a anchor point (block 148). Iterative processing of each cluster (block 149) and shared concept (block 150) continues, after which the routine returns.

FIGURE 12 is a routine for placing clusters 160 for use in the method of FIGURE 9. The purpose of this routine is to form cluster groupings or "groupers" of grafted cluster spines.

Each sublist of placeable clusters is iteratively processed (blocks 161-175), as follows. For each sublist (block 161), if the sublist includes an anchor point (block 162), the anchor point is selected (block 165). Otherwise, a new grouper is started (block 163) and the first cluster in the sublist is selected as the anchor point and removed from the sublist (block 164). Each cluster in the sublist is then iteratively processed (block 166-173), as follows.

For each cluster (block 166), the radius of the cluster is determined (block 167) and the routine attempts to place the cluster along the open vectors emanating from the anchor point (block 168). The radius is needed to ensure that the placed clusters do not overlap. If the cluster was not successfully placed (block 169), the cluster is skipped and processed during a further iteration (block 175). Otherwise, if the cluster is successfully placed (block 169) and is also designated as an anchor point (block 170), the angle of the anchor point is set (block 171), as further described below with reference to FIGURE 12. The cluster is then placed in the vector (block 172). Processing continues with the next cluster (block 173).

Upon the completion of the processing of each cluster in the sublist (block 166), the angle for the cluster is set if the cluster is selected as an anchor point for a grafted cluster (block 174). Processing continues with the next sublist (block 175), after which the routine returns.

FIGURE 13 is a graph 180 showing, by way of example, an anchor point within a cluster spine generated by the cluster display system of FIGURE 1. Anchor points 186, 187 are formed along an open edge at the intersection of a vector 183a, 183b, respectively, drawn from the center 182 of the cluster 181. The vectors are preferably drawn at a normalized angle, such as 60° in the described embodiment, relative to the vector 188 forming the cluster spine.

A cluster 181 functioning as an anchor point can have one or more open edges depending upon the placement of adjacent clusters and upon whether the cluster 181 is the largest endpoint, smallest endpoint or, as shown, midpoint cluster. In the described embodiment, largest endpoint clusters have four open edges, smallest endpoint clusters have three open edges, and midpoint clusters have two open edges. Adjusting the normalized angle and allowing more (or less) overlap between grafted cluster spines are possible to allow for denser (or sparser) cluster placements.

FIGURE 14 is a routine for placing groupers 190 in the use of the method of FIGURE 9. The purpose of this routine is to place groupings of cluster sublists into a visual display space.

Each of the groupers is iteratively processed (block 191-197), as follows. For each grouper (block 191), if the grouper comprises a singleton cluster (block 192), the grouper is skipped (block 197). Otherwise, if the grouper is the first grouper selected (block 193), the grouper is centered at the origin of the visual display space (block 194). Otherwise, the angle of the grouper and radius from the center of the display are incremented by the size of the grouper, plus extra space to account for the radius of the end-point cluster at which the cluster is grafted (block 195) until the grouper can be placed without substantially overlapping any previously-placed grouper. Slight overlap within 20° between clusters is allowed. A grouper is added to the display space (block 196). Iterative processing continues with the next grouper (block 197). Finally, all singleton groupers are placed in the display space (block 198). In the described embodiment, the singleton groupers are placed arbitrarily in the upper left-hand corner, although other placements of singleton groupers are possible, as would be recognized by one skilled in the art. The routine then returns.

Although the foregoing method 120 of FIGURE 8 has been described with reference to circular clusters, one skilled in the art would recognize that the operations can be equally applied to non-circular clusters forming closed convex volumes.

While the invention has been particularly shown and described as referenced to the embodiments thereof, those skilled in the art will understand that the foregoing and other changes in form and detail may be made therein without departing from the spirit and scope of the invention.